Support-set bottlenecks for video-text representation learning

Year: 2020 | Citations: 260 | Authors: Mandela Patrick, Po-Yao (Bernie) Huang, Yuki M. Asano, Florian Metze, Alexander Haupt

---

Abstract

The dominant paradigm for learning video-text representations -- noise contrastive learning -- increases the similarity of the representations of pairs of samples that are known to be related, such as text and video from the same sample, and pushes away the representations of all other pairs. We posit that this last behaviour is too strict, enforcing dissimilar representations even for samples that are semantically-related -- for example, visually similar videos or ones that share the same depicted action. In this paper, we propose a novel method that alleviates this by leveraging a generative model to naturally push these related samples together: each sample's caption must be reconstructed as a weighted combination of other support samples' visual representations. This simple idea ensures that representations are not overly-specialized to individual samples, are reusable across the dataset, and results in representations that explicitly encode semantics shared between samples, unlike noise contrastive learning. Our proposed method outperforms others by a large margin on MSR-VTT, VATEX and ActivityNet, for video-to-text and text-to-video retrieval.